SRNT

STIC-ILL

From: Goldberg, Jeanine
Sent: Tuesday, January 28, 2003 3:12 PM
To: STIC-ILL
Subject: please pull 5' cdna library

1. GENOMICS, (1996 Nov 1) 37 (3) 327-36.
   Journal code: 8800135. ISSN: 0888-7543.

2. DNA RESEARCH, (1997 Feb 28) 4 (1) 61-6.
   Journal code: 9423827. ISSN: 1340-2838.

3. GENE, (2001 Jan 24) 263 (1-2) 93-102.
   Journal code: 7706761. ISSN: 0378-1119.

THANK YOU

Jeanine Enewold Goldberg
1634
CM1--12D11
Mailbox-- 12E12
306-5817

ELSEVIER

Gene 263 (2001) 93–102

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Comparative evaluation of 5′-end-sequence quality of clones in CAP trapper and other full-length-cDNA libraries

Yuichi Sugahara[a], Piero Carninci[a,*], Masayoshi Itoh[a], Kazuhiro Shibata[a], Hideaki Konno[a,b], Toshinori Endo[a], Masami Muramatsu[a,b], Yoshihide Hayashizaki[a,b,1]

[a]Laboratory for Genome Exploration Research Project, Genomic Sciences Center and Genome Science Laboratory, RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan
[b]CREST, JST, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

Received 15 June 2000; rceived in revised form 2 October 2000; accepted 16 November 2000
Received by T. Gojobori

Abstract

To enhance the usefulness of the laboratory mouse and to facilitate the rapid assay of gene functions we have been collecting the entire set of mouse full-length cDNA by one-pass sequencing. To collect full-length cDNA clones efficiently, it is critical to construct high-quality cDNA libraries. In recent years, we have been developing a way to construct full-length cDNA libraries by using biotinylation of the cap structure (the 'CAP-trapper' method) coupled with treatment to increase reverse transcriptase efficiency at high temperature by the addition of trehalose. In this paper we report our evaluation of the quality of CAP trapper and a number of other full-length cDNA libraries, including the results of 5′ end analysis of clones in CAP trapper and the other libraries. We used a procedure that compared the 5′-ends of cDNA clones with those of genes in the public databases. Our analysis showed that 63% of cDNA clones in CAP trapper libraries had sequences that were either the same length as those of equivalent genes in the public database or 5′-extended, and that 90% of these clones maintained their coding sequences. These results indicate that the CAP trapper library is a promising tool for collecting full-length cDNA in large-scale projects. Comparison of the quality of CAP trapper with that of other full-length-cDNA libraries confirmed the value of these libraries. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Expressed sequence tags; cDNA Library; Trehalose; Transcriptome; cDNA Encyclopedia; Mammalian gene collection

1. Introduction

In studying genes the information on the coding sequence is essential. Although many attempts have been made to predict the transcription unit from genomic sequence data, the accuracy of these predictions is still limited. A more direct and efficient approach to gathering information on the coding sequences is to construct and sequence cDNA libraries. Several efforts for large-scale sequencing of cDNA libraries have been progressing, and most data are generated by single-pass sequencing of randomly selected cDNA clones: that is, by expressed-sequence-tag (EST) projects (Adams et al., 1991, 1995; Hillier et al., 1996; Marra et al., 1999). The EST data collection is huge, and ESTs are used in many genetic studies, including gene identification, expression studies and gene mapping. ESTs are also used in genomic sequencing projects to identify splicing sites and overlapped transcription units. However, the usefulness of EST clones is limited: because many EST clones lack the complete sequences of mRNAs, they cannot be used to reveal the primary structures of entire genes and encoded proteins. To investigate the biological function of a gene it is necessary to obtain the complete gene product: this corresponds to the full-length cDNA, which carries the complete protein coding sequence plus the untranslated regions (UTR). Although it is possible to obtain a full-length clone starting from ESTs, this requires several rounds of clone-screening of a library or cloning after primer exten-

Abbreviations: ESTs, expressed sequence tags; dbEST, ESTs database; 5′-SPSs, 5′ single pass sequences; KG, known genes; BLAST, basic local alignment search tool; CDS, coding sequence

[*] Corresponding author. RIKEN, Genome Science Laboratory, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan. Tel.: +81-298-369145; fax: +81-298-369098.

E-mail addresses: carninci@rtc.riken.go.jp, rgscerg@rtc.riken.go.jp (P. Carninci); yosihide@rtc.riken.go.jp, rgscerg@rtc.riken.go.jp (Y. Hayashizaki).

[1] Co-corresponding author, to whom questions related to the RIKEN cDNA Clones Bank should be addressed at RIKEN, Genome Science Laboratory, 3-1-1 Kayadai, Tsukuba, Ibaraki 305-0074, Japan. Tel.: +81-298-369145; fax: +81-298-369098.

0378-1119/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(00)00557-6

sion such as in the rapid amplification of cDNA ends (RACE). This work is laborious and unsuitable for whole genome-scale projects.

An alternative to gene discovery strategies based on ESTs followed by cloning of individual full-length cDNAs is to construct full-length-cDNA libraries and to sequence them. In recent years we have been making large-scale efforts to collect and sequence mouse full-length-cDNA clones (RIKEN Mouse cDNA Encyclopedia Project) for the majority of genes (http://genome.rtc.riken.go.jp). We have clustered already more than 128,500 cDNAs representing the majority of mouse genes. To do this it is essential to construct and evaluate high-quality cDNA libraries on a routine basis. Our aim is to use libraries that are of high complexity with low levels of contamination of truncated clones. They must be suitable for large-scale sequencing, and the majority must maintain their full-length sequences or at least their coding sequences to be included in the full-length-cDNA library. Traditionally, the construction of such libraries has been difficult for two technical problems. One problem is that it is difficult to synthesize full-length cDNAs efficiently, mainly because of the limited processivity of reverse transcriptase and the stops induced by the secondary structure of mRNAs, which cause detachment of the enzyme from the elongating complex. The other is that it is difficult to separate the full-length cDNAs from the incomplete ones, and to clone the full-length cDNAs selectively. Recently, several methods have been established to enable libraries to be enriched with full-length cDNAs. These methods are oligo-capping (Maruyama and Sugano, 1994; Kato et al., 1994), Capfinder (from Clontech, now called 'SMART') and CAPture (Edery et al., 1995). These methods use the cap, a structure present at the 5′-end of eukaryotic mRNA, to select full-length cDNAs.

We have been developing two techniques to address the above problems. The first is the use of trehalose-thermoactivated reverse transcriptase, which allows the preparation of longer cDNA at high temperatures (Carninci et al., 1998). The second is a 5′-end-selection technique that uses biotinylation of the cap structure. This technique is called CAP trapper (Carninci et al., 1996). At present, we are sequencing only CAP trapper libraries and we have produced more than 170 of them. When we construct a new library we pick up a limited number of clones and their suitability examined for large-scale sequencing. 5′-end sequencing is performed to evaluate the quality of the libraries. The 5′-end-sequence data are then computationally analyzed by comparing them with sequences in the public databases. In this report, we describe the evaluation of the 5′-end quality of the CAP trapper libraries made for our project and the transcriptional start of some genes. Deepness and redundancy of the CAP trapper libraries are described elsewhere (Carninci et al., 2000). We also compare CAP trapper libraries and other full-length libraries prepared by the oligo-capping and Capfinder methods. Large-scale cDNA sequences are not available from libraries produced with the fourth method, CAPture (Edery et al., 1995) that therefore could not be included in this analysis.

2. Materials and methods

2.1. Database sources

We obtained the database sequences to be used for comparison with our 5′ single pass sequences (5′-SPSs) from GenBank (available from 15 December 1998) at the National Center for Biotechnology Information (NCBI). We constructed the known gene database from gbrod.seq by collecting the entries for which the organism was '*Mus musculus*' and whose definitions included 'mRNA' and 'complete cds' (or 'complete CDS'). We also obtained the EST database from the NCBI anonymous ftp site (ftp://ncbi.nlm.nih.gov/blast/db/est_mouse). The EST database is a collection of rodent sequence data, and we used it unmodified; i.e. we did not select the *Mus musculus* data. Therefore, our cDNA data and those from organisms other than *Mus musculus*, such as the rat or *Mus spretus*, could not be compared.

2.2. Computer analysis of the 5′-ends of cDNA clones

We evaluated the quality of CAP trapper libraries by comparing the 5′-SPSs of each cDNA clone with that of the equivalent gene in the public database. We used at first the criterion that the 5′-ends of the cDNA clones should be the same length as, or more extended than, the sequence in the public database. Therefore, the frequency of occurrence of 5′-extended clones was one of the measures of library quality. The comparison was performed in two steps. First, we ran a Basic Local Alignment Search Tool (BLAST, version 1.4.10 MP) (Altschul et al., 1990) search. We used E value (Expectation value) $= 1 \times 10^{-8}$ for the known gene database and $E = 1 \times 10^{-20}$ for the EST database) to list rapidly the homologous sequences in the database that satisfied E-values lower than the above; these became candidates for homologous to cDNA clones. BLAST provided those database sequences that were locally homologous to the 5′-SPSs of the cDNA clones. Second, we calculated the global homology between the 5′-SPSs of the cDNA clones and the locally homologous database sequences by the Smith–Waterman algorithm. In the Smith–Waterman calculation, we limited the window size to within two base pairs (bp) in order to avoid large insertion or deletion sequence errors. The gap penalty was set as -1.0 for 1 bp insertion-deletions (indel), and -2.0 for 2 bp indel. If the 5′-SPS of a cDNA clone overlapped more than 50 bp with a candidate sequence and had a more-than-90% similarity in the overlapped region, we regarded the 5′-SPS and the database sequence as belonging to the same gene. We calculated the differences in the 5′-end sequences and evaluated how many base pairs of cDNA clones were extended or truncated. In the comparison with the known gene database, we identified some cDNA clones that were comparable to more than one known gene (KG) (or more than one database sequence due to the redundant presence of sequences of a given gene in the databases). In these cases, if the 5′-end of the cDNA clone was the same length as or longer than at least one of them, we categorized the cDNA clone as L_KG.

2.3. Source of sequence data for oligo-capping and Capfinder libraries

We obtained the sequence data for the three mouse oligo-capping libraries from the UniGene database at NCBI. For this study we used lib132, lib135 and lib136. According to the library information supplied, the source of lib132 was liver, for lib135 was embryo, and for lib136 was kidney.

The Capfinder library was constructed from mouse blastocysts and sequenced from 5′-ends in our laboratory (Sasaki et al., 1998).

3. Results and discussions

3.1. Statistics of CAP trapper libraries

We summarized the results of our comparison of the 5′-SPSs of cDNA clones and those of gene sequences from the known database (Table 1). We analyzed over 8000 sequences and identified more than 1000 clones with known genes, named 'hit'. Library variability can be seen in Table 1. Although the 5′-end selection process that we used (CAP trapping) is common, the final library quality may differ due to starting quality of mRNA and construction steps, including normalization, subtraction and size-fractionation. We categorized 63% of the 'hit' clones as long

Table 1
Quality of the CAP trapper library, as evaluated by comparing the 5′-end sequences of each cDNA clone with those of genes from the known gene databases

	Library	Clone[a]	Hit[b]	Hit L_KG[c] (%)	Hit L_CDS[d] (%)	Gene[e]	Gene L_KG[f] (%)	Gene L_CDS[g] (%)
6	Kidney	922	199	80	99	79	67	97
7	Brain	100	8	63	75	6	50	67
8	Lung	77	3	67	100	3	67	100
9	Spleen	377	56	21	98	11	82	100
10	Heart	909	139	65	94	85	69	94
11	Embryo-18	490	50	56	88	32	63	84
12	Lung	541	95	68	98	73	66	97
13	Liver	25	6	100	100	5	100	100
14	Brain	79	15	73	100	9	67	100
15	Cerebellum	157	31	74	90	19	63	84
16	Placenta	48	12	83	100	12	83	100
17	Testis	50	2	50	100	2	50	100
18	Pancreas	18	1	100	100	1	100	100
20	Small intestine	411	54	52	87	50	53	86
21	Liver-Lung mix	366	39	28	38	29	31	41
22	Stomach	452	28	64	82	26	65	83
23	Tongue	522	56	48	77	51	48	76
24	ES cell	316	44	80	89	29	75	86
25	Embryo-13 (liver)	600	77	43	92	44	55	91
26	Embryo-10	275	52	75	92	32	71	91
27	Embryo-11	259	31	52	87	28	54	86
28	Embryo-10 + Embryo-11	404	49	76	94	38	71	93
29	Hippocampus	130	10	60	70	8	56	69
30	Embryo-12 (head)	216	16	63	81	16	63	81
31	Embryo-13 (head)	223	29	79	86	26	77	85
32	Embryo-14 + Embryo-17 (head)	161	17	65	88	17	65	88
33	Embryo-17 (head)	96	8	75	88	8	75	88
34	Embryo-10	140	3	67	67	3	67	67
36	Brain	27	1	100	100	1	100	100
Total		8391	1131	63	90	445	57	84

[a] Number of clones successfully sequenced from the 5′-end.
[b] Number of clones corresponding to any known gene, redundantly.
[c] Number of clones whose 5′-ends were the same length as or longer than the corresponding known gene (KG) sequences.
[d] Number of clones that maintained at least the annotated first ATG.
[e] Number of different genes hit in the analysis of a given cDNA library (non-redundant number).
[f] Value, normalized per the number of genes (fair contribution), which 5′-ends were the same length as or longer than the corresponding known gene (KG) sequences (see text).
[g] Value, normalized per the number of genes (fair contribution), which maintained at least the annotated first ATG (see text).

known genes ('hit L_KG') clones, and 90% as long CDS ('hit L_CDS') clones.

3.2. Expression-normalized evaluation of library quality

Our evaluation of library quality was based on the ratio of L_KG and L_CDS based on 'hits' to any clone, either they represent a gene multiple times or only once (see Table 1). It is known that some genes are easily cloned as full-length forms and some are not. In addition, some genes are abundantly expressed in specific tissues or at specific stages of development. If a highly abundant gene were easily cloned as a full-length form, we would overestimate the quality of the library. To avoid such overestimation, we also evaluated the quality of libraries by treating the contribution of each gene fairly. We calculated the contribution of the full-length cDNA as if they would appear once (non-redundantly) in the analyzed cDNA libraries and called 'gene'. To do this, we first calculated the percentage of L_KG and L_CDS clones for each gene ('gene', 'gene L_KG' and 'gene L_CDS' columns in Table 1). For example, if there were four clones identified as one known gene, and three of them were longer than the database sequence and one was shorter, we would consider that 75% (3/4) of the clones were L_KG clones for this gene. We then performed the same analysis for all the genes and clones in a given library, and averaged the number of L_KG and L_CDS clone sequences for all the genes. Normalized for all genes, 57% of the clones were L_KG clones and 84% were L_CDS clones. These data are essential for the preparation of a non-redundant set of clones if one clone is to be randomly selected from each gene cluster. Since our purpose was to collect a set of cDNA clones that could be translated, 90% of the clones and 84% of the genes would be L_CDS (Table 1).

In the known gene database, several sequences partly or completely lack the 5′ UTR sequence. As well as comparing the 5′-SPSs of cDNA clones with the known gene database sequences, it would be advisable to evaluate the quality of the library against a set of independent data. Although the percentage of ESTs with full-length sequences is low, the very large number of ESTs makes it possible to find several 5′ UTR sequences in the EST database. Additionally, EST database sequences may include many more species of genes than the known gene database, making them potentially suitable for quality comparisons. Therefore, we compared the 5′-SPSs of cDNA clones and the EST database sequences (Table 2). Essentially, we used the same method of analysis used for the known gene database. In Table 2, 'L_EST' means the number of clones whose 5′-SPSs were the same length as or longer than the most 5′-extended EST sequences (5′-longest EST). The EST database was highly redundant, and when we compared it with the 5′-SPSs, we identified several more EST sequences than were found in known genes. Of these sequences, we used the one that had the most extended 5′-end. Therefore, for all practical purposes, 'L_EST' means that the clone was the same length as, or longer than, all the equivalent EST sequences. Against the EST database, 37% of overlapping CAP trapper cDNA clones showed the same or extended 5′-ends.

Our analysis showed that a large number of the CAP trapper library clones were of high quality, being categorized into L_KG, L_CDS or L_EST clones. However, the presence of incomplete cDNA clones in the CAP trapper library suggests that the procedure of library construction was not completely efficient in the selection of full-length cDNA or that there was some degradation of the cDNA before cloning. Alternatively, clones slightly shorter than databases could simply be isoforms transcribed from different promoters or different starting points. To estimate the nature and degree of the contamination of shorter clones, we examined the characteristics of those clones that were shorter than the L_KG and L_EST database equivalents (Fig. 1). We measured the degree of truncation by the number of base pairs (bp) of 5′-difference between the 5′-SPSs and the database sequences (GenBank and EST). If we defined the 'tentatively acceptable' clones as those fulfilling

Table 2
Quality of the CAP trapper library, as evaluated by comparing the 5′-end sequences of the cDNA clones and those of genes in the EST database

	Library	Clone[a]	Hit[b]	L_EST[c] (%)
6	Kidney	922	680	232 (34)
7	Brain	100	38	12 (32)
8	Lung	77	27	10 (37)
9	Spleen	377	285	106 (37)
10	Heart	909	569	245 (43)
11	Embryo-18	490	294	96 (33)
12	Lung	541	318	143 (45)
13	Liver	25	21	5 (24)
14	Brain	79	31	21 (68)
15	Cerebellum	157	128	42 (33)
16	Placenta	48	30	12 (40)
17	Testis	50	26	12 (46)
18	Pancreas	18	15	11 (73)
20	Small intestine	411	276	105 (38)
21	Liver-Lung mix	366	133	41 (31)
22	Stomach	452	226	106 (47)
23	Tongue	522	293	120 (41)
24	ES cell	316	210	82 (39)
25	Embryo-13 (liver)	600	432	155 (36)
26	Embryo-10	275	209	43 (21)
27	Embryo-11	259	155	56 (36)
28	Embryo-10 + Embryo-11	404	254	90 (35)
29	Hippocampus	130	73	34 (47)
30	Embryo-12 (head)	216	113	35 (31)
31	Embryo-13 (head)	223	148	44 (30)
32	Embryo-14 + Embryo-17 (head)	161	99	29 (29)
33	Embryo-17 (head)	96	62	32 (52)
34	Embryo-10	140	18	3 (17)
36	Brain	27	15	7 (47)
Total		8391	5178	1929 (37)

[a] Number of clones successfully sequenced from the 5′-end.

[b] Number of clones corresponding to a known gene.

[c] Number of clones whose 5′-SPSs were the same length as or longer than the most 5′-extended EST sequence (5′-longest EST).

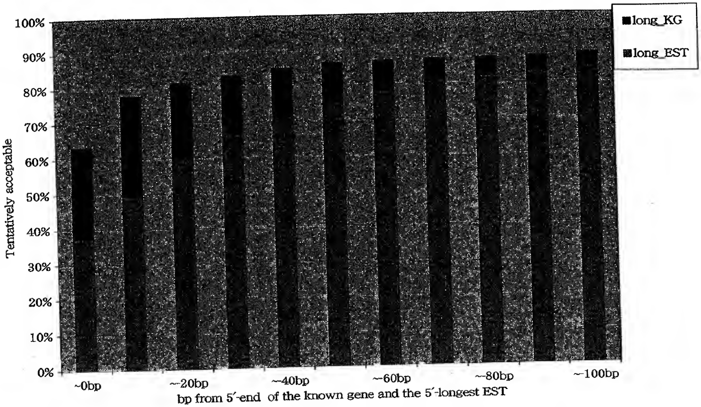

Fig. 1. Distribution of tentatively acceptable clones shorter than the L_KG and L_EST database equivalents. L_KG indicates the percentage of our clones longer than the longest entry in GenBank, while L__EST indicated the percentage of our clones longer than the longest EST.

the criterion of covering the 5′ end region of the longest sequence cloned, then the percentage of these clones was represented as class 0 bp on the vertical axis (see Fig. 1). When we relaxed this criterion, the percentage of 'tentatively acceptable' clones rapidly increased and reached a plateau within 100 bp. This implies that the length difference was small compared with the total mRNA length for most clones, which had nearly the same 5′-ends as the database sequences but for a certain percentage they were slightly shorter. The appearance of shorter clones might have been caused by cloning artifacts or by the presence of differential promoters/starting sites. In fact, some genes have multiple promoters or transcription starting-sites that are used alternatively depending on the expression level, the tissue and the developmental stage (Davis and Schultz, 1998). Therefore, a fraction of the clones that looked shorter might in fact be the full-length version in certain specific tissues. Additionally, the dbEST sequences might have included some 5′-end extra sequences in case of lack of vector trimming, as confirmed in some cased by manual inspection (not shown); this may have caused apparent truncation of our clones. For instance, if we changed the criterion to specify 'tentatively acceptable' clones as having 0–20 bp, then the L_KG content would increase from 63 to 82%. In a recent paper, oligo-capping libraries were evaluated in similar manner (Marra et al., 1999). In that paper, the shorter clones within 50 bp of the known gene database were defined as nearly full-length. Although our analysis was not consistent with others, if we were to adopt the same criteria, then 86% of the clones would be nearly full-length. Currently we are score our libraries using as criteria the percentage of clones which lengths are within 100 bp of the longest ESTs.

3.3. General 5′-end comparison between clones in CAP trapper libraries and other full-length libraries

The 5′-end comparison between the 5′-SPSs of the cDNA clones and the database sequences revealed that CAP trapper libraries included a large portion of clones that were L_KG or L_CDS. Currently, two other full-length library-construction methods are available, namely oligo-capping and Capfinder, while sequences produced with CAPture method, are not available. The 5′-end-selection techniques for these methods are different from those used in CAP trapper, and the library qualities differ (Table 3). All three methods have the ability to enrich the numbers of high-quality cDNA clones, being the average of full-length cDNA rate higher than the ~27% of full-length rate of ESTs (Marra et al., 1999) obtained mainly by sequencing conventional normalized/subtracted cDNA libraries (Bonaldo et al., 1996). In fact, whatever full-length cloning technique is used, we can expect that long cDNAs will be more difficult to clone as a full-length. In order to estimate

Table 3
Comparative quality of the CAP trapper, oligo-capping and Capfinder libraries, as evaluated by comparing the 5′-end sequences of clones with those of genes in the known gene database[a]

Libraries	Clone	Hit	L_KG (%)	L_CDS (%)
Cap-trapper	8391	1131	717 (63)	1022 (90)
Oligo-capping	7942	1899	941 (50)	1457 (77)
Capfinder	3212	732	637 (87)	691 (94)

[a] Notations as for Table 1.

the size bias in full-length cDNA cloning, we examined the correlation between the length of the gene and the L_CDS clones from the three technologies in terms of completeness of the coding region (Fig. 2). Here, the length of the gene was determined by the base pairs recorded in GenBank. Although the recorded sequence length may not have been accurate, we considered that we had truly represented the size of each gene. As expected, as genes became longer, the percentage of high clones with the first ATG decreased. However, the decrease was not dramatic, since more than 66% of cap-trapper clones still maintained the first ATG even when their starting mRNAs were longer than 4.0 Kb. There was an apparent drop in the number of clones longer than 5 Kb; this may be explained by a decrease in the ability of each method to clone longer cDNAs or possibly by instability of the long plasmid clones. Additionally, since we know very few of the 5′-end sequences for the large genes, we may have underestimated the quality of the long cDNA clones, because the 5′ end of any given long gene may not have been completely characterized. This would not be the case of shortened clones. This hypothesis is supported by the fact that in cDNA libraries that contain genes selected for large size there is a high rate of unknown sequences. For instance, more than 65% of the genes in the >7 Kb size-selected cDNA library #14 have unknown sequences against ESTs, whereas fewer than 10% of the cDNAs from other libraries used in this analysis had unknown sequences against ESTs (data not shown). CAP trapper appears superior to oligo-capping because of overall higher rate of capture of full-length cDNA, but inferior to Capfinder if the sizes of the cloned cDNAs are not taken into consideration (Table 3). In fact, although Capfinder could produce very good full-length rate for short cDNAs, it failed to produce any full-length cDNA for mRNAs longer than 3 Kb (Fig. 2), shoving that Capfinder is not optimal for long full-length due probably to the use of polymerase chain reaction (PCR). Additionally, Capfinder cDNA libraries showed, during the preparation of the full-length cDNA encyclopedia, a marked decrease in complexity if compared to cap-trapper cDNA libraries (P. Carninci, manuscript submitted), which would affect gene discovery if using

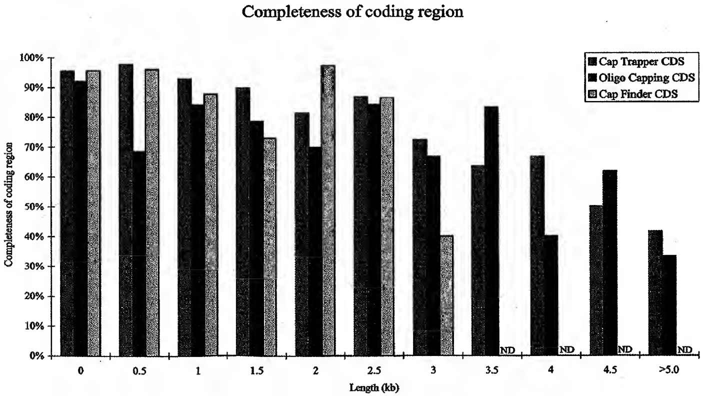

Fig. 2. Correlation between the lengths of genes in the known database and the L_CDS of clones from all three libraries in terms of completeness of the coding region. The abbreviation ND in correspondence of the Capfinder column indicated that no data is available (no clones corresponding to long mRNAs were found).

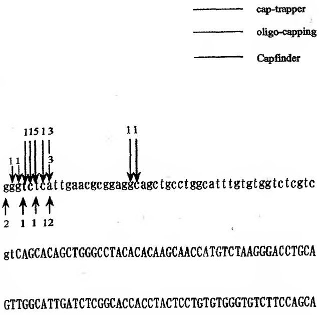
cap-trapper
oligo-capping
Capfinder
115 1 3
11
1 1
3
gggtctcattgaacgcggaggcagctgcctggcatttgtgtggtctcgtc
2 1 1 12
gtCAGCACAGCTGGGCCTACACACAAGCAACCATGTCTAAGGGACCTGCA
GTTGGCATTGATCTCGGCACCACCTACTCCTGTGTGGGTGTCTTCCAGCA

Fig. 3. Comparisons of the 5′-end sequence of the breast heat-shock protein gene with those of clones derived by the three different methods (CAP trapper, oligo-capping and Capfinder). The arrows indicate the 5′-ends of the clones, and the number associated with each arrow represents the number of clones with the 5′-ends indicated. The capital letters in the sequence correspond to the data recorded in GenBank (accession number gbU27129). The small letters represent new sequences obtained from our work. The red arrows correspond to the CAP trapper clones, the blue arrows to the oligo-capping clones, and the light blue arrows to the Capfinder clones. The pink ATG is the annotated first ATG.

these libraries. From Fig. 2, CAP trapper appeared superior to oligo-capping in terms of the average representation of cDNAs carrying the first ATG, and both oligo-capping and CAP trapper appeared better than Capfinder for cloning long, full-length clones.

3.4. Detailed analysis of transcription starting sites

To determine a starting site for transcription some laborious experimental work, such as S1 mapping or RACE (Frohman et al., 1988), is required. This work is not easy to scale up in genomic-scale approaches. If we merely examine the sequencing information we cannot be sure that the most 5′-extended clone reflects the real starting site of transcription. In addition, some genes have more than one starting site for transcription due to the use of differential promoter or transcriptional starting points and some are alternatively spliced, so that we cannot be sure whether or not the shorter clones are full-length. In this situation, some might suggest that if clones that have exactly the same 5′-ends and are derived from full-length cDNA libraries appear frequently, they will be full-length. The following is the analysis of some few available sequences of cDNAs from the same gene that were obtained from libraries prepared with all or al least two of the methodologies discussed in this work. This analysis highlights the difficulties in the evaluation of full-lengthness of cDNAs and shows the degree the variability of the transcriptional starting point.

The methods used to construct full-length cDNA libraries select the 5′-end of mRNAs. It is possible that variations in 5′-end-selection techniques inherently lead to the preferential appearance of different 5′-ends of cDNA clones due to alternative transcription start or promoters or, alternatively, production of library-specific cloning artifacts. We therefore compared the 5′-ends of cDNA clones derived from the CAP trapper, oligo-capping and Capfinder libraries. Many clones of the breast heat-shock protein gene (Fig. 3) had almost the same 5′-ends, and the differences were limited to within 8 bp. The first few G residues might have been artifacts induced by reverse transcriptase, which is adds a C in a template-free fashion at the end of the first-strand cDNA. From these data there was no clear differences in the position of the 5′-ends or the frequency of clones with a certain 5′-end. This suggests that for the categorization of successful clones, the appearance of the 5′-end is independent of the library-construction method used.

In contrast with the case of breast heat-shock protein, the 5′-ends of cDNA clones of the ferritin light chain gene showed clear differences among the various library-construction methods (Fig. 4). The 5′-ends of the CAP trapper clones were located in the most 5′-upstream region (from 1–7 bp in Fig. 4). There were various source tissues for the CAP trapper clones: 13 clones came from the kidney, two from the spleen, two from stage 18 embryos, one from

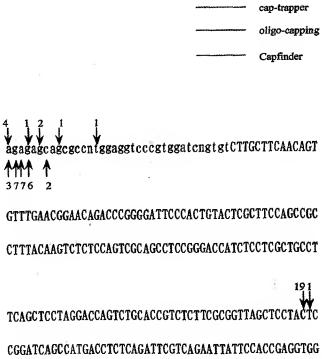
cap-trapper
oligo-capping
Capfinder
4 1 2 1 1
agagagcagcgccnggaggtcccgtggatcngtgtCTTGCTTCAACAGT
3776 2
GTTTGAACGGAACAGACCCGGGGATTCCCACTGTACTCGCTTCCAGCCGC
CTTTACAAGTCTCTCCAGTCGCAGCCTCCGGGACCATCTCCTCGCTGCCT
191
TCAGCTCCTAGGACCAGTCTGCACCGTCTCTTCGCGGTTAGCTCCTACTC
CGGATCAGCCATGACCTCTCAGATTCGTCAGAATTATTCCACCGAGGTGG

Fig. 4. Comparisons of the 5′-end sequence of the ferritin light chain (gbJ04716) with those of clones derived by the three different methods (CAP trapper, oligo-capping and Capfinder). Notations as for Fig. 3.

the lung, one from the placenta, three from ES (embryonic stem) cells, one from a mix of a stage 10 embryo and stage 11 embryo, one from the head of a stage 13 embryo and one from the brain. In contrast, the 5′-ends of the most dominant clones in the oligo-capping libraries were located at 198 bp. These clones were derived from three libraries with various source tissues: 14 clones came from the liver, two from embryos and three from the kidney. There were three oligo-capping clones (other than the most dominant ones) with 5′-ends located at 4, 15 or 199 bp; they were expressed in the liver. These differences might be caused by tissue specificity. However, 13 CAP trapper clones and three oligo-capping clones were derived from same source tissue (kidney). In addition, none of the CAP trapper clones had 5′-ends at 198 bp, and only one oligo-capping clone had a 5′-end in the most 5′-upstream region. The 5′-ends of the Capfinder clones were located in nearly the same region as those of the CAP trapper clones. These differences could indicate that the appearance of the clones at 198 bp was caused by treatment of RNA ligase or some sequence specific bias of the Oligo-capping method that was not observed with the CAP trapper and Capfinder methods. Alternatively, a step common to the CAP trapper and Capfinder methods could have caused the amplification or the preferential selection for cDNA around the 5′-upstream region. These data suggest that each library-construction method may have its own biases but do not clearly define what is the method that has the lowest bias.

Since the CAP trapper and oligo-capping methods select the cap structure of mRNA, the cDNA would, in principle, be full-length, although experimentally the selection is never complete. It is therefore very difficult to judge whether or not a clone is full length. A complete copy of each mature mRNA should be present no matter which library construction technique is used and the frequency of appearance of clones with the same 5′-ends may not be the proper criteria to assess whether or not a clone is full-length. Among data available from our and others databases, we have found some ambiguous cases (Fig. 5); the sequences of the apolipoprotein A-II gene and those of the corresponding cDNA clones were aligned. In this case, the two CAP trapper clones had the same 5′-ends. However, according to GenBank reports an annotated first ATG is located in the 5′-upstream region, so that these two clones would have no potential to express protein properly, if the protein sequence is the one reported in literature.

These two clones were derived from two different libraries (source tissues: ES cells and the liver of a stage 13 embryo); therefore they could not be sister clones, and had to be independent clones. In addition, the 5′-ends of these clones corresponded to the terminal region of exon 2; this may eliminate the possibility that the transcript was produced by alternative splicing. In another example (Fig. 6), we compared *cdc42* and two corresponding oligo-capping clones. In this case, seven clones (including these two) from three oligo-capping libraries (two clones from the liver, four from embryos and one from the kidney) had the same 5′-ends. However, the 5′-ends of these clones were located downstream of the first ATG, suggesting that the clones were truncated. As was the case with the ferritin light chain gene (see Fig. 4), no clone with the same 5′-end as those of the truncated CAP trapper clones appeared in the oligo-capping libraries for the apolipoprotein A-II gene, and no clone with same 5′-end as those of the truncated oligo-capping clones appeared in the CAP trapper libraries for *cdc42*. This fact implies that frequency of appearance of clones with the same 5′-ends does not necessarily mean that the clones are full-length, since the frequency of appearance of the clone may be biased by the library-construction method. We would like to suggest that if several clones have the same 5′-ends and they are derived from libraries constructed with different 5′-end-enrichment methods, then they are strong candidate to be full-length. In the case of the breast heat-shock protein gene (see Fig. 3), the 5′-ends of the clones were almost the same, strongly indicates that the clones were full-length. However, this may not be possible for many clones because of different complexity of libraries prepared in this analysis, and further developments will include scoring of features of mRNAs and comparison with genomic sequence.

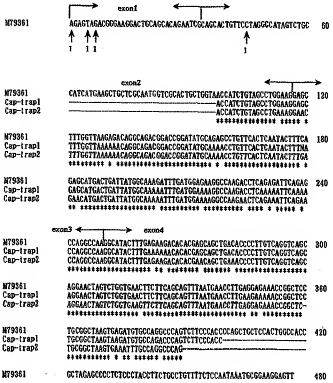
exon1
M79361 AGAGTAGACGGGAAGGACTGCAGCACAGAATCGCAGCACTGTTCCTAGGCCATAGTCTGC 60
1 1 1 1
exon2
M79361 CATCATGAAGCTGCTCGCAATGGTCGCACTGCTGGTAACCATCTGTAGCCTGGAAGGAGC 120
Cap-trap1 ACCATCTGTAGCCTGGAAGGAGC
Cap-trap2 ACCATCTGTAGCCTGAAAGGAAC
M79361 TTTGGTTAAGAGACAGGCAGACGGACCGGATATGCAGAGCCTGTTCACTCAATACTTTCA 180
Cap-trap1 TTTGGTTAAAAAACAGGCAGACGGACCGGATATGCAAAACCTGTTCACTCAATACTTTNA
Cap-trap2 TTTGGTTAAAAAACAGGCAGACGGACCGGATATGCAAAACCTGTTCACTCAATACTTTGA
M79361 GAGCATGACTGATTATGGCAAAGATTTGATGGAGAAGGCCAAGACCTCAGAGATTCAGAG 240
Cap-trap1 GAGCATGACTGATTATGGCAAAAATTTGATGGAAAAGGCCAAGACCTCAAAAATTCAAAA
Cap-trap2 GAACATGACTGATTATGGCAAAAATTTGATGGAAAAGGCCAAGAACTCAGAAATTCAGAA
exon3
exon4
M79361 CCAGGCCAAGGCATACTTTGAGAAGACACACGAGCAGCTGACACCCCTTGTCAGGTCAGC 300
Cap-trap1 CCAGGCCAAGGCATACTTTGAAAAAACACACGAGCAGCTGACACCCCTTGTCAGGTCAGC
Cap-trap2 CCAGGCCAAGGCATACTTTGAGAAGACACACGAACAGCTGAAACCCCTTGTCAGGTCAGC
M79361 AGGAACTAGTCTGGTGAACTTCTTCAGCAGTTTAATGAACCTTGAGGAGAAACCGGCTCC 360
Cap-trap1 AGGAACTAGTCTGGTGAACTTCTTCAGCAGTTTAATGAACCTTGAAGAAAAACCGGCTCC
Cap-trap2 AGGAACTAGTCTGGTGAAGTTCTTCAGCAGTTTAATGAACCTTGAGGAGAAACCGGCTC-
M79361 TGCGGCTAAGTGAGATGTGCCAGGCCCAGTCTTCCCACCCCAGCTGCTCCACTGGCCACC 420
Cap-trap1 TGCGGCTAAGTAAGATGTGCCAGACCCAGTCTTCCCACC
Cap-trap2 TGCGGCTAAGTGAAATTTGCCAGGCCCAG
M79361 GCTAGAGCCCCTCTCCCTACCTTCTGCCTGTTTTCTCCAATAAATGCGGAAGGAGTT 480

Fig. 5. Sequence comparison between apolipoprotein A-II (gbM79361) and two CAP trapper clones. The arrows indicate other CAP trapper (red) and oligo-capping (blue) clones and their 5′-ends. The numbers of associated clones are indicated by arrows. The red ATG is the annotated first ATG. Other ATGs in the coding frame are indicated in pink. The stop codon is blue.

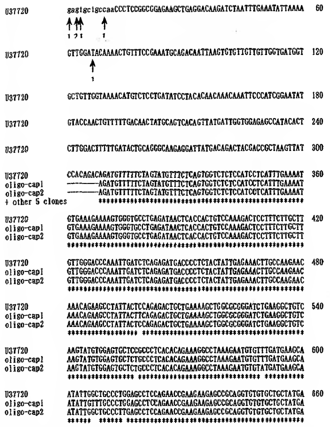
U37720 gagtgctgccaaCCCTCCGGCGGAGAAGCTGAGGACAAGATCTAATTTGAAATATTAAAA 60
1 2 1 1
U37720 GTTGGATACAAAACTGTTTCCGAAATGCAGACAATTAAGTGTGTTGTTGTTGGTGATGGT 120
1
U37720 GCTGTTGGTAAAACATGTCTCCTGATATCCTACACAACAAACAAATTCCCATCGGAATAT 180
U37720 GTACCAACTGTTTTTGACAACTATGCAGTCACAGTTATGATTGGTGGAGAGCCATACACT 240
U37720 CTTGGACTTTTTGATACTGCAGGGCAAGAGGATTATGACAGACTACGACCGCTAAGTTAT 300
U37720 CCACAGACAGATGTTTTTCTAGTATGTTTCTCAGTGGTCTCTCCATCCTCATTTGAAAAT 360
oligo-cap1 AGATGTTTTTCTAGTATGTTTCTCAGTGGTCTCTCCATCCTCATTTGAAAAT
oligo-cap2 AGATGTTTTTCTAGTATGTTTCTCAGTGGTCTCTCCATCCTCATTTGAAAAT
+ other 5 clones
U37720 GTGAAAGAAAAGTGGGTGCCTGAGATAACTCACCACTGTCCAAAGACTCCTTTCTTGCTT 420
oligo-cap1 GTGAAAGAAAAGTGGGTGCCTGAGATAACTCACCACTGTCCAAAGACTCCTTTCTTGCTT
oligo-cap2 GTGAAAGAAAAGTGGGTGCCTGAGATAACTCACCACTGTCCAAAGACTCCTTTCTTGCTT
U37720 GTTGGGACCCAAATTGATCTCAGAGATGACCCCTCTACTATTGAGAAACTTGCCAAGAAC 480
oligo-cap1 GTTGGGACCCAAATTGATCTCAGAGATGACCCCTCTACTATTGAGAAACTTGCCAAGAAC
oligo-cap2 GTTGGGACCCAAATTGATCTCAGAGATGACCCCTCTACTATTGAGAAACTTGCCAAGAAC
U37720 AAACAGAAGCCTATTACTCCAGAGACTGCTGAAAAGCTGGCGCGGATCTGAAGGCTGTC 540
oligo-cap1 AAACAGAAGCCTATTACTTCAGAGACTGCTGAAAAGCTGGCGCGGATCTGAAGGCTGTC
oligo-cap2 AAACAGAAGCCTATTACTCCAGAGACTGCTGAAAAGCTGGCGCGGATCTGAAGGCTGTC
U37720 AAGTATGTGGAGTGCTCCGCCCTCACACAGAAAGGCCTAAAGAATGTGTTTGATGAAGCA 600
oligo-cap1 AAGTATGTGGAGTGCTCTGCCCTCACACAGAAAGGCCTAAAGAATGTGTTTGATGAAGCA
oligo-cap2 AAGTATGTGGAGTGCTCTGCCCTCACACAGAAAGGCCTAAAGAATGTGTATGATGAAGCA
U37720 ATATTGGCTGCCCTGGAGCCTCCAGAACCGAAGAAGAGCCGCAGGTGTGTGCTGCTATGA 660
oligo-cap1 ATATTGTTTGCCCTGGAGCCTCCAGAACCGAAGAAGAGCCGCAGGTGTGTGCTGCTATGA
oligo-cap2 ATATTGGCTGCCCTTGAGCCTCCAGAACCGAAGAAGAGCCGCAGGTGTGTGCTGCTATGA

Fig. 6. Sequence comparison between *cdc42* (gbU37720) and two oligo-capping clones. Five other oligo-capping clones are present, although their sequence data are not shown in the figure. The small letters represent novel sequences discovered in this work. The other notations are as for Fig. 5.

4. Conclusions

We are using a large-scale approach to collect the majority of mouse full-length cDNAs. The key to success is to prepare high-quality libraries in which most of the clones are full-length. At the same time, library-construction methods should permit new genes to be discovered at a high rate. Therefore, we have discarded both those strategies that involve PCR, because they cause bias in the representation of clones (which appear less frequently with long mRNAs), and those strategies that involve RNA ligase, because they show sequence-specificity of the ligation reaction, thus causing skewed representation of cDNAs in the cDNA library. The CAP trapper strategy selectively captures 5′-end of mRNA and enables full-length cDNA to be enriched. Additionally, to improve complexity, most CAP trapper libraries use size-fractionation or normalization/subtraction. Complexity of cap trapper cDNA libraries is described elsewhere (P. Carninci et al., 2000), and on the average shows 2 or 3-fold higher gene discovery than both Oligo-capping and Capfinder cDNA libraries.

A number of other factors need to be considered in assessing the results of our analysis of library-construction methods. With all of the methods, the extensive gene manipulation that is required to facilitate the discovery of new genes may partly damage the quality of full-length cDNA. Various sources were used for the mRNAs used to prepare the cDNA libraries, and this factor cannot be taken easily into account when different libraries are being compared. Also, in our comparison, except where specified earlier, we used data obtained from the routine preparation of cDNA libraries as they appeared in our database (http://genome.rtc.riken.go.jp/). These data averages the values of those cDNA libraries that we do not consider to be excellent in terms of full-length genes. If we were presenting data only from the best libraries, the resulting library could be closer to the perfect full-length library. Since the 5′-end-sequence comparison of the cDNA clones and the public databases showed that 90% of the CAP trapper clones maintained at least the first ATG, we consider that our technology is excellent for both full-length cloning and routine new-gene discovery. In fact, we have been able to cluster more 128,500 clones which we estimate to represent 70 ~ 85,000 mouse genes (P. Carninci, submitted). When the CAP trapper libraries were compared with the other full-length cDNA libraries, the CAP trapper library appeared superior to other methods in terms of both the ability to clone long cDNAs and the numbers of full-length clones present. In addition, CAP trapper does not require PCR and/or RNA ligase; this permits unbiased cloning of long and rare cDNAs. Generally, the ability of all cloning technology to clone efficiently long cDNAs when mixed with short cDNAs is still problematic. Also, propagating plasmids containing long cDNAs is poses serious problems; this suggests that improved cloning systems, currently under development in our laboratory, will allow discovering new long, full-length cDNA clones more efficiently. Despite the current limitations, our analyses strongly suggest that the CAP trapper library is the most promising tool presently available for efficiently collecting full-length cDNA clones via an EST-like sequencing approach.

Acknowledgements

This study has been supported by: Special Coordination Funds and a Research Grant for the Genome Exploration Research Project from the Science and Technology Agency of the Japanese Government; a Grant-in-Aid for Scientific Research on Priority Areas and the Human Genome Program from the Ministry of Education and Culture of Japan; and a Grant-in-Aid for a Second Term Comprehensive 10-Year Strategy for Cancer Control from the Ministry of Health and Welfare of Japan to Y.H.

References

Adams, D.M., Kelly, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R., Venter, J.C., 1991. Complementary

DNA sequencing: expressed sequence tags and human genome project. Science 252, 1651–1656.

Adams, M.D., et al., 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377(Suppl.), 3–175.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 5, 403–410.

Bonaldo, M.F., Lennon, G., Soares, M.B., 1996. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 6, 791–806.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y., Schneider, C., 1996. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. Genomics 37, 327–336.

Carninci, P., Nishiyama, Y., Westover, A., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., Hayashizaki, Y., 1998. Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. Proc. Natl. Acad. Sci. USA 95, 520–524.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M., Hayashizaki, Y., 2000. Normalization and subtraction of Cap-Trapper selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. Genome Res. 10, 1607–1630.

Edery, I., Chu, L.L., Sonenberg, N., Pelletier, J., 1995. An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). Mol. Cell. Biol. 15, 3363–3371.

Davis Jr, W., Schultz, R.M., 1998. Molecular cloning and expression of the mouse translation initiation factor eIF-1A. Nucleic Acids Res. 26, 4739–4747.

Frohman, M.A., Dush, M.K., Martin, G.R., 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. USA 85, 8998–9002.

Hillier, L.D., et al., 1996. Generation and analysis of 280,000 human expressed sequence tags. Genome Res. 6, 807–828.

Marra, M., et al., 1999. An encyclopedia of mouse genes. Nat. Genet. 21, 191–194.

Maruyama, K., Sugano, S., 1994. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene 138, 171–174.

Kato, S., Sekine, S., Oh, S.W., Kim, N.S., Umezawa, Y., Abe, N., Yokoyama-Kobayashi, M., Aoki, T., 1994. Construction of a human full-length cDNA bank. Gene 150, 243–250.

Sasaki, N., Nagaoka, S., Itoh, M., Izawa, M., Konno, H., Carninci, P., Yoshiki, A., Kusakabe, M., Moriuchi, T., Muramatsu, M., Okazaki, Y., Hayashizaki, Y., 1998. Characterization of gene expression in mouse blastocyst using single-pass sequencing of 3995 clones. Genomics 49, 167–179.